Published in final edited form as:
Methods Mol Biol. 2014 ; 1137: 119–130. doi:10.1007/978-1-4939-0366-5_9.

SPOT-Seq-RNA: Predicting protein-RNA complex structure and RNA-binding function by fold recognition and binding affinity prediction

Yuedong Yang[a], **Huiying Zhao**[a], **Jihua Wang**[b], and **Yaoqi Zhou**[a]

[a]School of Informatics, Indiana University Purdue University, Indianapolis, Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, 719 Indiana Ave Ste 319, Walker Plaza Building, Indianapolis, Indiana 46202, USA.

[b]Shandong Provincial Key Laboratory of Functional Macromolecular Biophysics and Department of Physics Dezhou University, Dezhou 253023, China

Summary

RNA-binding proteins (RBPs) play key roles in RNA metabolism and post-transcriptional regulation. Computational methods have been developed separately for prediction of RNA-binding proteins and RNA-binding residues by machine learning techniques and prediction of protein-RNA complex structures by rigid or semi-flexible structure-to-structure docking. Here, we describe a template-based technique called SPOT-Seq-RNA that integrates prediction of RNA-binding proteins, RNA-binding residues, and protein-RNA complex structures into a single package. This integration is achieved by combining template-based structure-prediction software, SPARKS X, with binding-affinity prediction software, DRNA. This tool yields reasonable sensitivity (46%) and high precision (84%) for an independent test set of 215 RBPs and 5766 non-RBPs. SPOT-Seq-RNA is computationally efficient for genome-scale prediction of RNA-binding proteins and protein-RNA complex structures. Its application to human genome study has revealed a similar sensitivity and ability to uncover hundreds of novel RBPs beyond simple homology. The online server and downloadable version of SPOT-Seq-RNA are available at http://sparks.informatics.iupui.edu/server/SPOT-Seq-RNA/

[1]The query sequence must be a protein sequence in FASTA format. The gene in the DNA/RNA sequence has to be converted to amino acids sequence first. Unknown amino acids (e.g. X) must be removed.

[3]Here unaligned residues and the side-chain atoms except C_β atoms are excluded for interaction calculations so that we can prevent large fluctuation in predicted binding affinities due to possible atomic clashes between RNA and modeled side-chains or modeled missing residues. A new version is in progress to relax modeled side-chain and missing residues so that we can estimate the protein-RNA binding affinity based on all interactions between protein and RNA molecules.

[6]The running time depends on the size of the query protein. For the example given here (1zbiB, 136 residues), it takes 22 minutes on an Intel Pentium 4 3.4GHz, in which 14 minutes are due to PSI-BLAST.

[7]For online service, the results can be obtained from the webpage directly or from email if an email address is given. To save computing resources, please do not submit query sequences more than once. The status of your job can be found by clicking the link "Check the current Queue to prevent DUPLICATE submits" on the main webpage. The result of your job will only be kept for one month after completion.

[8]The recent blast package can be downloaded from: ftp://ftp.ncbi.nlm.nih.gov/blast/executables/release/LATEST/. Select the appropriate executable version for your system (e.g. blast-2.2.26-x64-linux.tar.gz for 64 bit Linux). The NR database can be downloaded from ftp://ftp.ncbi.nih.gov/blast/db/nr.XX.tar.gz. As of Feb. 23, 2013, XX includes 10 numbers from 00 to 09, and each file is about 700 megabytes.

1 Introduction

The majority of the human genome is coded for RNA transcripts. Only tiny fractions of these RNA transcripts are messenger RNAs that code for proteins. All RNA transcripts, most with unknown functions, are regulated by RNA-binding proteins from birth (transcription) to death (degradation). Thus, locating all RNA-binding proteins (RBPs) in a genome and determining protein-RNA complex structures are key steps for understanding the mechanism of post-transcriptional regulation and for mapping the network of protein-RNA interactions.

It is difficult to locate RBPs and determine their protein-RNA complex structures experimentally due to high flexibility of RNA structures and the difficulty associated with crystallization of complex structures. Despite this difficulty, there is a steady increase in the number of protein-RNA complex structures deposited in the protein data bank from 45 in 2001 to 180 in 2011 (non-redundant at 90% sequence identity or less) (1). Moreover, hundreds of novel, unconventional, or moonlighting RBPs have been discovered (2–4). Experimental discovery of new RBPs and determination of protein-RNA complex structures, however, is costly and inefficient. There is a need for the development of highly accurate bioinformatics tools for predicting RNA binding function and protein-RNA complex structures.

Most methods developed for predicting RNA-binding proteins are based on machine-learning methods that employ information of protein sequences and/or known protein structures (5,6). Meanwhile, docking techniques for protein-RNA interactions have been developed by using a scoring/energy function for protein-RNA interaction (7–10). Here, we describe SPOT-Seq-RNA, a template-based technique that combines predictions of protein-RNA complex structure and binding affinity (11). More specifically, SPOT-Seq-RNA employs a template library of non-redundant protein-RNA complex structures and attempts to match the query sequence to the protein structures in protein-RNA complexes by the fold recognition technique SPARKS X (12). Significant matches will be employed to predict the complex structures between a target sequence and template RNA as well as the binding affinity of the complex.

In SPOT-Seq-RNA, structure prediction is performed by the latest version of our fold recognition technique SPARKS X (12) which was among the best performing single automatic servers in several critical assessment of structure prediction (CASP) meetings (CASP 6 (13), CASP 7 (14) and CASP 9 (12)). SPARKS X is a multi-dimensional probabilistic matching between sequence profiles generated from PSI-BLAST (15) for query and template sequences and between structural features of a template and those predicted by SPINE X (16–18) for a query sequence. Predicted structural features include secondary structure (17), backbone torsion angles (16), and residue solvent accessibility (18). For binding affinity prediction, we extracted a knowledge-based energy function, DRNA, from protein-RNA complex structures (19) based on a distance-scaled finite ideal-gas reference (DFIRE) state (20). The DFIRE reference state was found to be one of the best reference states for deriving knowledge-based energy functions for folding and binding studies (21, 22). While many template-based structure prediction methods and knowledge-based energy functions for protein-RNA interactions exist, the coupling between fold recognition by SPARKS X and binding affinity prediction by DRNA in SPOT-Seq-RNA provides the first dedicated high-resolution function prediction for RBPs.

SPOT-Seq-RNA was cross-validated by leave-homology-out and independently tested by several datasets (11). It was found to significantly improve over a sequence-to-profile search technique, PSI-BLAST (15), and a profile-to-profile search technique, HHPRED (23), in

discriminating RBPs from non-RBPs. It was also shown to be far more sensitive and accurate in detecting RBPs than machine-learning based techniques, while having similar accuracy to the best machine-learning techniques for RNA-binding site prediction (24). More importantly, SPOT-Seq-RNA can provide a reasonably accurate prediction of protein-RNA complex structure (77% predicted structures having root-mean-squared distance of 4Å or less) (11). More recently, SPOT-Seq-RNA was applied to the human genome and independently tested by mRNA-binding proteins from a proteomic experiment (25). Discovery of more than 2000 novel RBPs in the human genome and validation of the results in messenger-RBPs by the proteomic experiment (4) confirm the usefulness of SPOT-Seq-RNA in predicting novel RNA-binding proteins beyond simple sequence homology and modeling of their complex structures.

2 Materials

2.1 Software

A software package is downloadable from our homepage with a shortcut link: http://sparks.informatics.iupui.edu/yueyang/download/index.php?Download=SPOT-Seq-RNA.tbz. This package as shown in Fig. 1 integrates one external program PSI-BLAST (15) and three in-house-built programs: SPINE X (structural property prediction) (16,17), SPARKS X (template-based structure prediction) (12), and DRNA (binding affinity prediction) (19).

1. An external program, PSI-BLAST, and protein NR database (15) are employed to generate a position-specific scoring matrix (PSSM) or sequence profile that is a required input for programs SPINE X and SPARKS X (see Note 2 to skip this step if a PSSM file is pre-calculated for the query sequence).
2. An in-house-made program, SPINE X (16–18), is applied to predict the secondary structure, torsional angles (ϕ and ψ), and the solvent accessible surface area (ASA). SPINE X is a neural-network predictor that couples secondary structure prediction with predictions of solvent accessibility and backbone torsion angles in an iterative manner. SPINE X was tested with a dataset of 2640 proteins and achieved an 82.0% accuracy in secondary structure prediction based on 10-fold cross validation. SPINE X can also be downloaded separately from our homepage.
3. SPARKS X is a template-based structure-prediction program. The program is employed to search for the best match between a query sequence and a template structure in the template database of protein-RNA complex structures. The statistically significant alignments from the best match (or matches) are utilized to construct complex structure models between the query and RNA of the template.
4. DRNA scoring function is used to calculate binding affinity. DRNA is a statistical energy function extracted from 174 protein-RNA complex structures with a distance-scaled finite ideal-gas reference state (19). It predicts the binding affinity based on the complex structure model between the query and template RNA.

2.2 Databases for RNA-binding proteins and non-RNA binding proteins

1. A prebuilt list of 1052 RNA-binding domains and chains and 5766 non-RNA binding chains were prepared. The files for template structural profiles for both RBPs and non-RBPs are located in the directory “TPL_input”. Here the database of RBPs contains template proteins in complex with their binding RNAs while all

[2]The package requires PSSM from PSI-BLAST. If a pre-calculated PSSM was prepared, PSI-BLAST can be skipped to save time. The user can choose to input a pre-calculated PSSM with option “-pssm” for the locally installed version.

non-RBPs serve as background statistics to calculate Z-scores to measure the significance of the matching template.

2. For RBPs the structural coordinates in PDB format were provided for model building. The coordinate files contain 1052 protein chains/domains as well as 632 RNAs from respective protein-RNA complexes, stored in directories “domains” and “RNA0”, respectively. Each protein file contains one chain or domain, while RNA c
3. oordinate files contain all RNA chains in the protein complex. These protein coordinate files are employed as the templates to build the structure model of the query protein while the RNA coordinates will be directly copied (with the same orientation as in the template complex structure) as the RNA conformation in the complex structure model for the query protein.

3 Methods

To describe the automated prediction pipeline as shown in Fig. 1, we used a protein, Bacillus halodurans RNase H catalytic domain mutant D132N (PDB id: chain B of 1zbi) as an example. This protein is RNase H and belongs to a nucleotidyl transferase superfamily, which includes transposase, retroviral integrase, Holliday junction resolvase, and RISC nuclease Argonaute (26).

3.1 Input and PSSM generation

The only input required for SPOT-Seq-RNA is the query protein sequence in FASTA format. Fig. 2A displays the input window for the web-based server that allows the cut-and-paste of the protein sequence RNase H. File upload is also allowed. Only one sequence per run is allowed for input. This sequence is subsequently passed to PSI-BLAST (15) to search for homologous sequences of the query sequence and to generate the position-specific substitution matrix (PSSM), which is constructed by three iterations of searching (E value less than 0.001) against the non-redundant (NR) sequence database.

3.2 Structural profile preparation for SPARKS X

The PSSM file (either given or generated from PSI-BLAST) above is first employed by SPINE X to predict protein structural properties, including secondary structure (in three states), torsional angles (ϕ and ψ), and the solvent accessible area (ASA), along with their respective confidence scores. SPINE X is a neural-network predictor that utilizes a Perl script file to automatically call five separate predictors that were compiled by the Intel Fortran compiler. Structural properties predicted by SPINE X together with PSSM are stored in a profile file (pro.inp) as input for SPARKS X. In this profile file, the first line indicates the residue number (NRES) of the query protein. The second line contains the sequence in one-letter code. The next 20 lines are the inverse value of PSSM for 20 amino acid residue types at all sequence positions (20×NRES). These 20 lines are arranged alphabetically based on residue names (ACDE...Y). Here, the inverse value of PSSM is used to reduce the number of characters in the file as we noticed that most values in the PSSM are negative. After that, another 20 lines are the probability of 20 amino acids at each sequence position. Then, the next four lines are predicted probabilities of secondary structures (three states of coil, helix, and sheet, CHE), ϕ, ψ, and relative ASA. These structural properties are followed by predicted confidence scores for secondary structure, ϕ and ψ, respectively. In the current version, the confidence score for ASA is pre-calculated based on amino-acid residue types.

3.3 Scanning over all templates by SPARKS X

The above sequence and structural profiles for the query sequence are employed by SPARKS X to compare with corresponding profiles of all template structures in the template library (see 2.2). The raw profile-profile alignment score in SPARKS X is calculated as follows:

$$\begin{aligned} S(i,j) = & -\frac{1}{200}[F^{\mathrm{seq}}_{\mathrm{query}}(i)\cdot M^{\mathrm{seq}}_{\mathrm{template}}(j) + F^{\mathrm{seq}}_{\mathrm{template}}(j)\cdot M^{\mathrm{seq}}_{\mathrm{query}}(i)] \\ & + w_1 E\left(SS_t(i)\middle| SS_q(j), C_{SS,q}(j)\right) \\ & + \sum_{k=2}^{4} w_k E\left(\Delta^k_{ij}\middle| C_{k,q}(j)\right) + S_{\mathrm{shift}} \end{aligned} \quad (1)$$

where w_k are weight parameters and S_{shift} is a constant. The first term in Eq. (1) is the profile-profile comparison between the sequence profile from the query sequence and that from the template sequence, where *F* and *M* are sequence-derived frequency profile and log odd profile, respectively. The second term is the energy term based on probabilistic matching between predicted secondary structures of the query and actual secondary structures of the template: $E\left(SS_t(i)\middle| SS_q(j), C_{SS,q}(j)\right) = -\ln\left(\frac{P(SS_t|SS_q, C_{SS,q})}{P(SS_t)}\right)$, where $P(SS_t|SS_q, C_{SS,q})$is the probability of the predicted secondary structure SS_q by SPINE X with confidence score $C_{SS,q}$ for a native secondary structure SS_t. Similarly, the next three terms are the energy terms based on probabilistic matching between other structural properties: $E(\Delta^k\middle| C_{k,q}) = -\ln\left(\frac{P(\Delta^k|C_{k,q})}{P^0(\Delta^k|C_{k,q})}\right)$, where $P(\Delta^k|C_{k,q})$ is the probability of the difference Δ^k between the predicted properties and corresponding native values with a confidence score of $C_{k,q}$. The reference probability $P^0(\Delta^k|C_{k,q})$ is obtained by comparing the predicted values to all native values in a dataset as described below. There are a total of three terms with k = 2 for real-value ϕ value, k = 3 for real-value ψ value, and k = 4 for real-value solvent accessibility. All energy terms were obtained from a non-redundant data set of 2479 proteins with length less than 500 amino acids from the original SPINE database [25% sequence identity cutoff, X-ray resolution of 3Å or higher, and no unknown structural regions] (27).

The raw alignment scores optimized by dynamic programming techniques for all templates are saved in the file called “pro.out”, in which each line contains the template name, the raw alignment score, the total alignment length including gaps, the number of gaps in two termini, the start and end positions of the query chain segment with effective alignment, and the number of exactly matched residue types in the alignment. The number of gaps can be positive (gaps in the query protein) or negative (gaps in the template protein). The sequence position begins counting from zero.

3.4 Selecting statistically significant matching templates

From the alignment raw score, the *Z-score* was calculated based on a normalized score $S_{norm} = S_{raw}/L^{\alpha}$ using the standard definition: *Z-score* = $(S_{norm} - S_{ave})/\Delta S$, where S_{raw} and *L* are the raw alignment score and alignment length (i.e. the second and third column in the pro.out file); α is 0.75; and S_{ave} and ΔS are the average value and standard error of the normalized score on all templates. A higher Z-score indicates a highly significant matching template from the average templates. Based on our previous statistics, templates with Z-scores of six or higher have 90% probability of having the same structural fold as the query protein.

By default, the program will record five or more templates with the highest Z-score or Z-scores greater than eight in file “pro.zs1”. In the file, the first column is the calculated Z-score followed by the query protein name and raw alignment scores that are output by SPARKS X for each template. These templates will be subjected to model building and binding affinity evaluation.

3.5 Building and evaluating protein-RNA complex models

All top matching templates in the file “pro.zs1” are used to build complex models. The model structures are built based on the alignment between the query and template sequences (see Note 5). The coordinates of the main-chain and C_β atoms (if present) of residues in the template will be copied to the corresponding aligned residues in the query protein. If the C_β of a residue except GLY is missing in the template, the C_β atom will be built based on the coordinates of the three main-chain heavy atoms (N, C_α, C). For those query residues not aligned to template residues, they will be ignored. The final protein model copies the RNA structure from the template to produce the complex structure model. All complex structure models are saved in separate files in PDB format (e.g. a complex built using template “2qk9A” will be saved in file “pro_2qk9A.pdb”).

From these complex structure models, the binding free energy will be evaluated by using the program DRNA. The pairwise distance-dependent energy between an atom of an amino acid residue and an atom in a RNA base is

$$u_{ij}^{DRNA}=\begin{cases} -\eta\ln\dfrac{N_{\rm obs}(i,j,r)}{\left(\frac{f_i^v(r)f_j^v(r)}{f_i^v(r_{\rm cut})f_j^v(r_{\rm cut})}\right)^{\beta}(\frac{r}{r_{\rm cut}})^{\alpha}N_{\rm obs}(i,j,r_{\rm cut})}, r<r_{\rm cut} \\ 0, r\geq r_{\rm cut} \end{cases}$$

where α=1.61, β=0.5, r_{cut} is the interaction cutoff distance (15Å), and the volume-fraction factor $f_i^v(r)=\sum_j N_{\rm obs}^{\rm protein-RNA}(i,j,r)/\sum_j N_{\rm obs}^{\rm All}(i,j,r)$, $N_{\rm obs}(i,j,r)$ is the number of observed pairs of atoms i and j at a given distance *r* from a database of protein-RNA complex structures. We employed residue/base-specific atom types with a total of 253 atom types (167 for protein and 86 for RNA). We also set the factor η arbitrarily to 0.01 to control the magnitude of the energy score. The statistics of N_{obs}(i,j,r) is saved in the file “dfire_RNA”. The binding free energy of a complex structure model is obtained by summing the interactions between any RNA atoms and protein atoms of main-chain atoms and C_β only with a distance less than 6.0Å. The calculated binding free energy together with the Z-score for all complexes is recorded in the file “pro.zs_en”.

3.6 Detecting RNA-binding proteins

The query protein is an RNA binding protein when both Z-score and energy thresholds are satisfied for at least one complex structure model. The final output file, “pro.result”, contains the template name, Z-score and the estimated binding free energy of the protein-RNA complex structure. The complex structure is then further employed to predict residues that interact with RNA (binding residue prediction). The binding residues were defined if any atom of the residue is less than 4.5Å to any RNA atoms. For a balance of coverage and accuracy of the prediction, we have set a threshold of 8.04 and −0.565 for Z-score and binding free energy, respectively. These thresholds were obtained in our benchmark studies

[5]SPARKS X has the option “-print level” to control different levels of outputs. The default level (0) only outputs the alignment score and length information. This will reduce the size of the output file in the template scanning step. A level equal to or greater than two will also print out the alignment between query and templates.

by maximizing the Matthews correlation coefficient (MCC) for two-state prediction of RNA-binding proteins (11).

If not a single template is found to satisfy both thresholds, the query protein will be considered as a non-RNA binding protein. However, we continue to present the top five matched templates and predicted complex structure models because low sensitivity (about 46%) may incorrectly predict some RBPs as non-RBPs despite correct prediction of complex structures. Users may have additional biological information to judge correctness of the complex structure model and function prediction (see Note 4).

3.7 SPOT-Seq-RNA input/output

Fig. 2 shows the input and output windows of the SPOT-Seq-RNA server at http://sparks.informatics.iupui.edu/server/SPOT-Seq-RNA/. This output is based on the query protein Bacillus halodurans RNase H. There are four matching templates within both Z-score and binding thresholds. Thus, this protein is predicted as an RNA-binding protein. Predicted complex structural models are listed according to respective templates. After filtering homologous templates, 2qk9A (17.6% sequence identity to the query sequence) was selected to demonstrate the overall accuracy of prediction. Fig. 3 displays the structurally aligned predicted and native complex structures by SPalign (28). One hundred and seven residues out of 136 residues (79%) in the query protein are aligned with its actual native structure with RMSD 2.8Å. In addition to the four templates within the thresholds, there is a template “1hysA0” that satisfied the Z-score but not the binding threshold. This illustrates the possibility of false negatives despite accurate structure prediction (see Note 4).

Acknowledgments

Funding for this work was supported by the National Institutes of Health grants [GM R01 085003 and GM R01 067168 (Co-PI) to Y.Z] and by the National Natural Science Foundation of China [grant 61271378 to J.W.].

References

1. Bernstein FC, Koetzle TF, Williams GJ, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. The Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol. 1977; 112:535–542. [PubMed: 875032]
2. Tsvetanova NG, Klass DM, Salzman J, Brown PO. Proteome-wide search reveals unexpected RNA-binding proteins in Saccharomyces cerevisiae. PLoS One. 2010; 5
3. Scherrer T, Mittal N, Janga SC, Gerber AP. A Screen for RNA-Binding Proteins in Yeast Indicates Dual Functions for Many Enzymes. PLoS One. 2010; 5:e15499. [PubMed: 21124907]
4. Castello A, Fischer B, Eichelbaum K, Horos R, Beckmann BM, Strein C, Davey NE, Humphreys DT, Preiss T, Steinmetz LM, et al. Insights into RNA Biology from an Atlas of Mammalian mRNA-Binding Proteins. Cell. 2012; 149:1393–1406. [PubMed: 22658674]
5. Puton T, Kozlowski L, Tuszynska I, Rother K, Bujnicki JM. Computational methods for prediction of protein-RNA interactions. J Struct Biol. 2012 in press.
6. Walia RR, Caragea C, Lewis BA, Towfic FG, Terribilini M, El-Manzalawy Y, Dobbs D, Honavar V. Protein-RNA interface residue prediction using machine learning: an assessment of the state of the art. BMC Bioinformatics. 2012; 13:89. [PubMed: 22574904]
7. Perez-Cano L, Solernou A, Pons C, Fernandez-Recio J. Structural prediction of protein-RNA interaction by computational docking with propensity-based statistical potentials. Pac Symp Biocomput. 2010; 15:269–280.

[4]Some predicted non-RNA binding proteins within the boundary of thresholds may be false negatives and have correctly predicted binding models. The strict cutoffs in Z-score and binding affinity were determined to maximize the MCC in our benchmark (low sensitivity around 46% but high precision at 84%) (11). For those templates with a Z-score greater than six but less than eight, the model protein structure is likely correct.

8. Zheng S, Robertson TA, Varani G. A knowledge-based potential function predicts the specificity and relative binding energy of RNA-binding proteins. Febs Journal. 2007; 274:6378–6391. [PubMed: 18005254]
9. Tuszynska I, Bujnicki JM. DARS-RNP and QUASI-RNP: new statistical potentials for protein-RNA docking. BMC Bioinformatics. 2011; 12:348. [PubMed: 21851628]
10. Setny P, Zacharias M. A coarse-grained force field for Protein-RNA docking. Nucleic acids research. 2011; 39:9118–9129. [PubMed: 21846771]
11. Zhao H, Yang Y, Zhou Y. Highly accurate and high-resolution function prediction of RNA binding proteins by fold recognition and binding affinity prediction. RNA Biology. 2011; 8:988–996. [PubMed: 21955494]
12. Yang Y, Faraggi E, Zhao H, Zhou Y. Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of the query and corresponding native properties of templates. Bioinformatics. 2011; 27:2076–2082. [PubMed: 21666270]
13. Zhou HY, Zhou Y. SPARKS 2 and SP3 Servers in CASP 6. Proteins. 2005; 61:152–156. [PubMed: 16187357]
14. Liu S, Zhang C, Liang SD, Zhou Y. Fold recognition by concurrent use of solvent accessibility and residue depth. Proteins. 2007; 68:636–645. [PubMed: 17510969]
15. Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research. 1997; 25:3389–3402. [PubMed: 9254694]
16. Faraggi E, Yang YD, Zhang SS, Zhou Y. Predicting Continuous Local Structure and the Effect of Its Substitution for Secondary Structure in Fragment-Free Protein Structure Prediction. Structure. 2009; 17:1515–1527. [PubMed: 19913486]
17. Faraggi E, Zhang T, Yang Y, Kurgan L, Zhou Y. SPINE X: Improving protein secondary structure prediction by multi-step learning coupled with prediction of solvent accessible surface area and backbone torsion angles. J. Computational Chemistry. 2011; 33:259–263.
18. Faraggi E, Xue B, Zhou Y. Improving the prediction accuracy of residue solvent accessibility and real-value backbone torsion angles of proteins by guided-learning through a two-layer neural network. Proteins. 2009; 74:847–856. [PubMed: 18704931]
19. Zhao HY, Yang YD, Zhou YQ. Structure-based prediction of RNA-binding domains and RNA-binding sites and application to structural genomics targets. Nucleic Acids Research. 2011; 39:3017–3025. [PubMed: 21183467]
20. Zhou HY, Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci. 2002; 11:2714–2726. [PubMed: 12381853]
21. Zhou Y, Zhou HY, Zhang C, Liu S. What is a desirable statistical energy function for proteins and how can it be obtained? Cell Biochem Biophys. 2006; 46:165–174. [PubMed: 17012757]
22. Zhou YQ, Duan Y, Yang YD, Faraggi E, Lei HX. Trends in template/fragment-free protein structure prediction. Theor Chem Acc. 2011; 128:3–16. [PubMed: 21423322]
23. Soding J, Biegert A, Lupas AN. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Research. 2005; 33:W244–W248. [PubMed: 15980461]
24. Zhao H, Yang Y, Zhou Y. Prediction of RNA binding proteins comes of age from low resolution to high resolution. Submitted. 2013
25. Zhao H, Yang Y, Janga SC, Kao C, Zhou Y. Charting the unexplored RNA-binding protein atlas of the human genome by combining structure and binding predictions. Submitted. 2013
26. Nowotny M, Gaidamakov SA, Crouch RJ, Yang W. Crystal structures of RNase H bound to an RNA/DNA hybrid: substrate specificity and metal-dependent catalysis. Cell. 2005; 121:1005–1016. [PubMed: 15989951]
27. Dor O, Zhou Y. Achieving 80% ten-fold cross-validated accuracy for secondary structure prediction by large-scale training. Proteins. 2007; 66:838–845. [PubMed: 17177203]
28. Yang Y, Zhan J, Zhao H, Zhou Y. A new size-independent score for pairwise protein structure alignment and its application to structure classification and nucleic-acid binding prediction. Proteins. 2012; 80:2080–2088. [PubMed: 22522696]

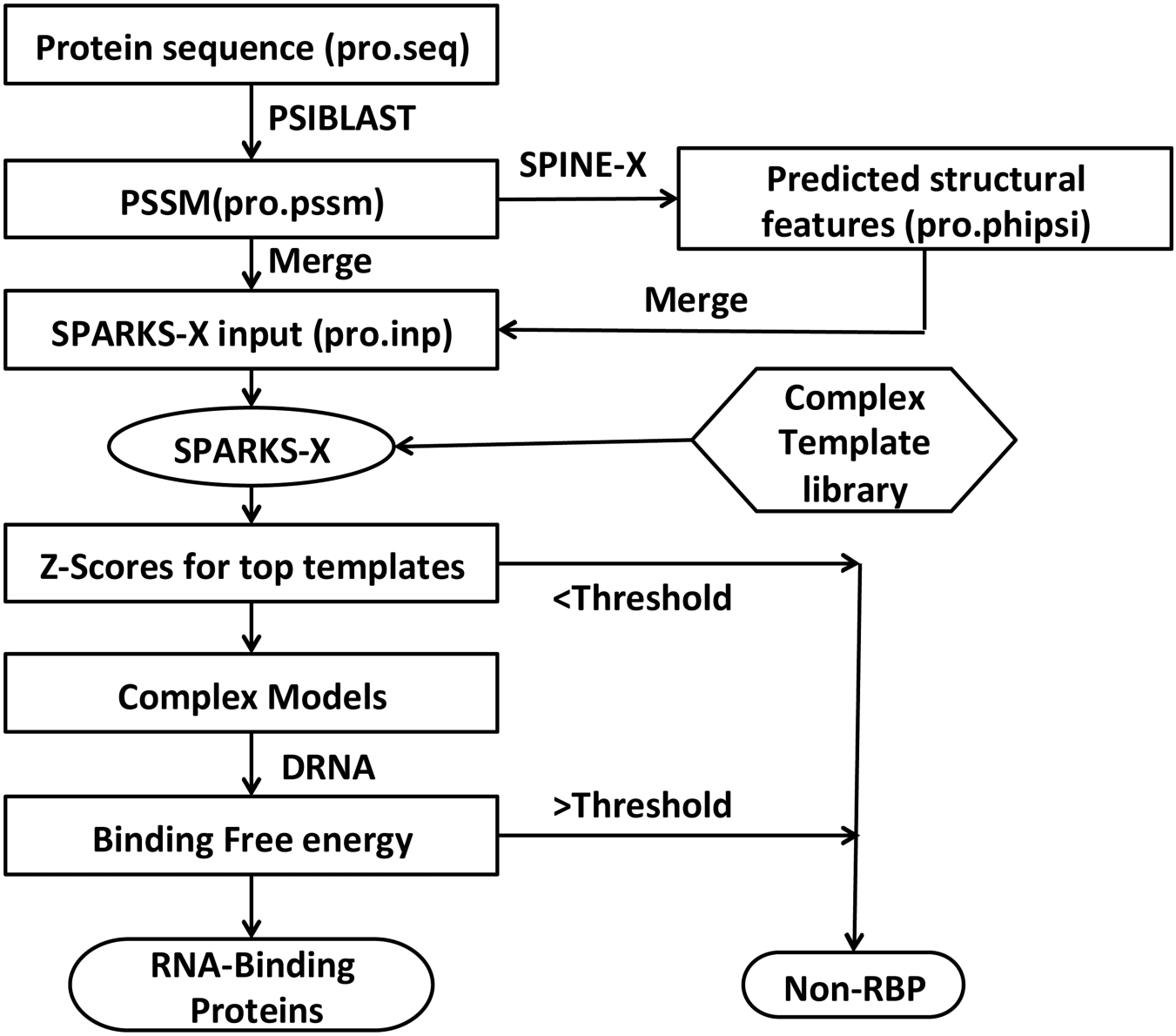

Figure 1.
The flow chart of SPOT-Seq-RNA.

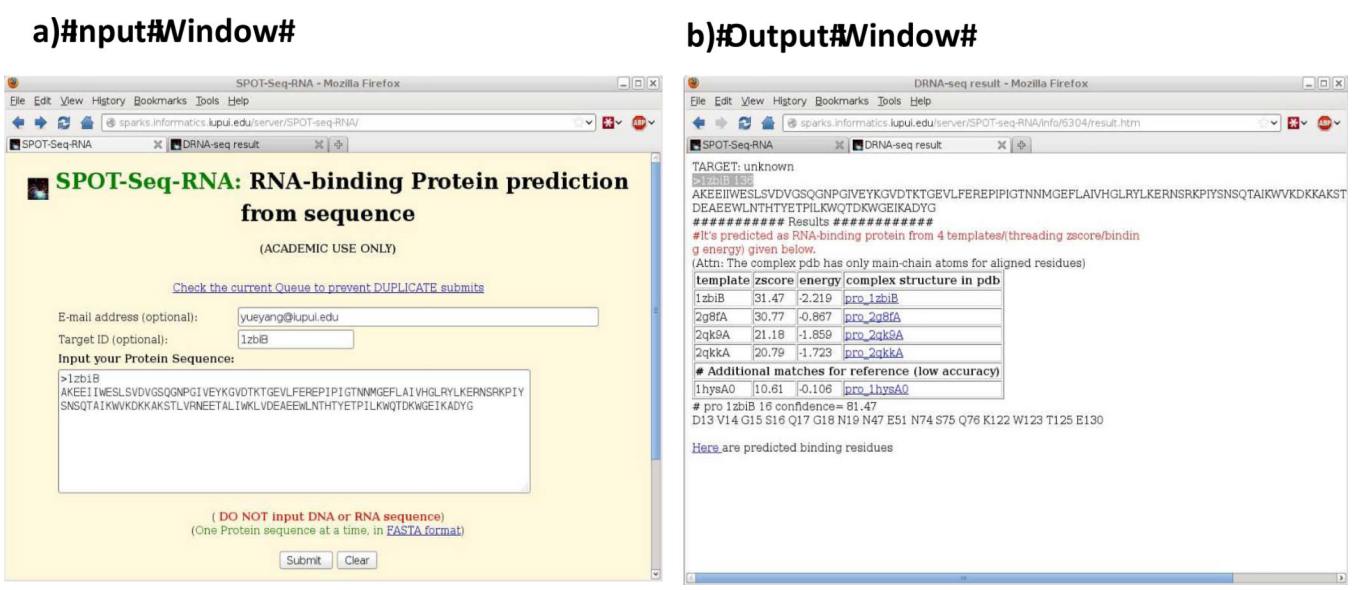

Figure 2.

The input and result windows of the SPOT-Seq-RNA server for the query protein Bacillus halodurans RNase H (PDBID: 1zbiB).

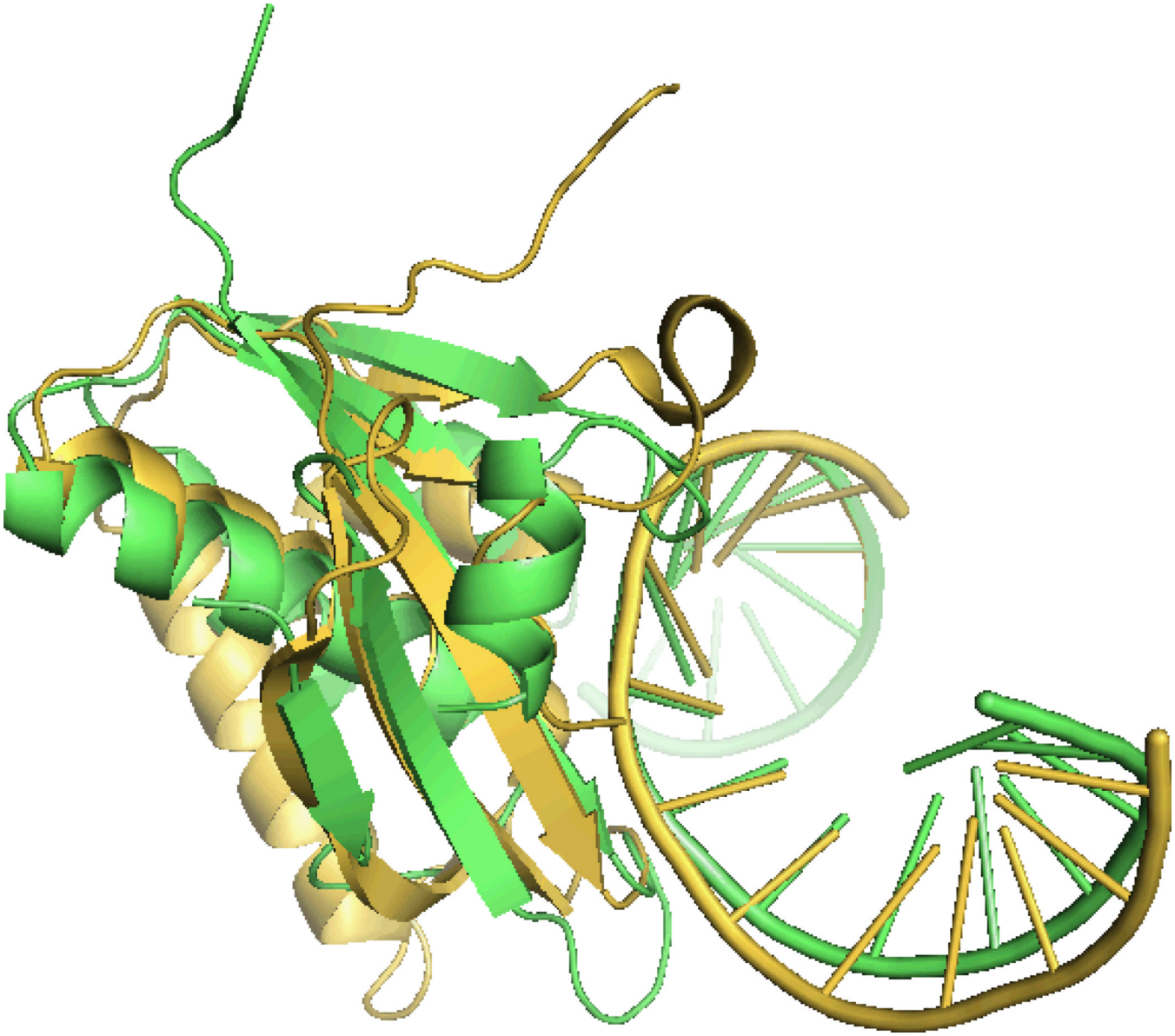

Figure 3.
The predicted model based on template 2qk9A for protein Bacillus halodurans RNase H (colored in green) is structurally aligned by SPalign to the native structure (colored in yellow). One complementary DNA chain has been removed.